A Survey on Reinforcement Learning

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Reinforcement Learning (RL) has emerged as a significant research area in the field of artificial intelligence, with a wide range of applications in robotics, finance, healthcare, and gaming Ngan Le (2021). The primary goal of RL is to develop algorithms that allow agents to learn optimal policies through interaction with their environment, maximizing the cumulative reward over time Kai Arulkumaran (2017). Despite the considerable progress made in recent years, RL still faces several challenges, such as the trade-off between exploration and exploitation, the curse of dimensionality, and the need for efficient algorithms that can handle large-scale and complex problems Sergey Ivanov (2019).

One of the major breakthroughs in RL has been the development of Q-learning algorithms, which have been proven to converge to the optimal solution Barber (2023). However, Q-learning is known to suffer from overestimation bias, leading to suboptimal performance and slow convergence in some cases Li Meng (2021). To address this issue, researchers have proposed various modifications and extensions to Q-learning, such as Double Q-learning Ehud Lehrer (2015) and Self-correcting Q-learning Rong Zhu (2020), which aim to mitigate the overestimation bias while maintaining convergence guarantees.

Another essential aspect of RL research is the incorporation of deep learning techniques, giving rise to the field of Deep Reinforcement Learning (DRL) Mahipal Jadeja (2017). DRL has demonstrated remarkable success in various domains, such as playing video games directly from pixels and learning control policies for robots Kai Arulkumaran (2017). However, DRL algorithms often require a large amount of data and computational resources, which limits their applicability in real-world scenarios Sergey Ivanov (2019). To overcome these limitations, researchers have proposed various approaches, including distributed DRL Qiyue Yin (2022) and expert-guided DRL Li Meng (2021), which aim to improve the sample efficiency and scalability of DRL algorithms.

Related work in the field of RL has also focused on the development of policy gradient methods, which optimize the policy directly by following the gradient of the expected return Ehsan Imani (2018). These methods have been particularly successful in continuous action settings and have led to the development of algorithms such as Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) van Heeswijk (2022). However, policy gradient methods often require on-policy data, which can be inefficient in terms of sample complexity Kmmerer (2019).

In summary, this survey aims to provide a comprehensive overview of the current state of Reinforcement Learning, focusing on the challenges and recent advances in Q-learning, Deep Reinforcement Learning, and policy gradient methods. By examining the key algorithms, techniques, and applications in these areas, we hope to shed light on the current limitations and future research directions in the field of RL.

2 RELATED WORKS

Markov Decision Processes: The study of reinforcement learning is fundamentally rooted in the understanding of Markov decision processes (MDPs). A concise description of stochastic approximation algorithms in reinforcement learning of MDPs is provided by Krishnamurthy (2015). The work done in Ehud Lehrer (2015) offers a full characterization of the set of value functions of MDPs, while Philip S. Thomas (2015) specifies a notation for MDPs. The concept of decisiveness in denumerable Markov chains has been extended to MDPs in Nathalie Bertrand (2020), exploring the implications of resolving non-determinism in adversarial or cooperative ways. Additionally, Arie Leizarowitz (2007) introduces an embedding technique to produce a finite-state MDP from a countable-state MDP, which can be used as an approximation for computational purposes.

Q-Learning and Variants: Q-learning is a widely used reinforcement learning algorithm that converges to the optimal solution Barber (2023). However, it is known to overestimate values and spend too much time exploring unhelpful states. Double Q-learning, a convergent alternative, mitigates some of these overestimation issues but may lead to slower convergence Barber (2023). To address the maximization bias in Q-learning, Rong Zhu (2020) introduces a self-correcting algorithm that balances the overestimation of conventional Q-learning and the underestimation of Double Q-learning. This self-correcting Q-learning algorithm is shown to be more accurate and achieves faster convergence in certain domains.

Expert Q-Learning: Expert Q-learning is a novel deep reinforcement learning algorithm proposed in Li Meng (2021). Inspired by Dueling Q-learning, it incorporates semi-supervised learning into reinforcement learning by splitting Q-values into state values and action advantages. An expert network is designed in addition to the Q-network, which updates each time following the regular offline minibatch update. The algorithm is demonstrated to be more resistant to overestimation bias and achieves more robust performance compared to the baseline Q-learning algorithm.

Policy Gradient Methods: Policy gradient methods are widely used for control in reinforcement learning, particularly in continuous action settings. Natural gradients have been extensively studied within the context of natural gradient actor-critic algorithms and deterministic policy gradients van Heeswijk (2022). The work in Ehsan Imani (2018) presents the first off-policy policy gradient theorem using emphatic weightings and develops a new actor-critic algorithm called Actor Critic with Emphatic weightings (ACE) that approximates the simplified gradients provided by the theorem. This algorithm is shown to outperform previous off-policy policy gradient methods, such as OffPAC and DPG, in finding the optimal solution.

Deep Reinforcement Learning: Deep reinforcement learning (DRL) combines the power of deep learning with reinforcement learning, achieving remarkable success in various domains, such as finance, medicine, healthcare, video games, robotics, and computer vision Ngan Le (2021). The field has seen significant advancements in recent years, with central algorithms such as the deep Q-network, trust region policy optimization, and asynchronous advantage actor-critic being developed Kai Arulkumaran (2017). A detailed review of DRL algorithms and their theoretical justifications, practical limitations, and empirical properties can be found in Sergey Ivanov (2019).

Temporal Networks: Temporal networks, where links change over time, are essential in understanding the ordering and causality of interactions between nodes in various applications. The work in Xiu-Xiu Zhan (2021) proposes a temporal dissimilarity measure for temporal network comparison based on the fastest arrival distance distribution and spectral entropy-based Jensen-Shannon divergence. This measure is shown to effectively discriminate diverse temporal networks with different structures and functional distinctions.

In conclusion, reinforcement learning has seen significant advancements in recent years, with various algorithms and techniques being developed to address the challenges in the field. From understanding the fundamentals of MDPs to developing advanced DRL algorithms, researchers continue to push the boundaries of what is possible in reinforcement learning and its applications.

3 BACKGROUNDS

3.1 PROBLEM STATEMENT AND FOUNDATIONAL CONCEPTS

Reinforcement Learning (RL) is a subfield of machine learning that focuses on training agents to make decisions in an environment to maximize a cumulative reward signal. In RL, an agent interacts with an environment through a sequence of actions, observations, and rewards, aiming to learn an optimal policy that maps states to actions Philip S. Thomas (2015). The problem can be formalized as a Markov Decision Process (MDP), which is defined by a tuple (S, A, P, R, γ), where S is the set of states, A is the set of actions, P is the state transition probability function, R is the reward function, and γ is the discount factor Ehud Lehrer (2015). The goal of RL is to find a policy $\pi(a|s)$ that maximizes the expected cumulative reward, defined as $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$, where R_{t+k+1} is the reward received at time step $t + k + 1$ Krishnamurthy (2015).

3.2 Q-Learning and Related Algorithms

Q-learning is a popular model-free RL algorithm that estimates the action-value function $Q(s, a)$, which represents the expected cumulative reward of taking action a in state s and following the optimal policy thereafter Barber (2023). The Q-learning update rule is given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[R(s,a) + \gamma \max_{a'} Q(s', a') - Q(s,a) \right],$$

where α is the learning rate, $R(s, a)$ is the reward for taking action a in state s, and s' is the next state Barber (2023). However, Q-learning can suffer from overestimation bias, which can lead to suboptimal performance Li Meng (2021). To address this issue, Double Q-learning was proposed, which uses two separate Q-value estimators and updates them alternately, mitigating overestimation bias while maintaining convergence guarantees Barber (2023). Another variant, Expert Q-learning, incorporates semi-supervised learning by splitting Q-values into state values and action advantages, and using an expert network to assess the value of states Li Meng (2021).

3.3 Policy Gradient Methods

Policy gradient methods are another class of RL algorithms that optimize the policy directly by estimating the gradient of the expected cumulative reward with respect to the policy parameters Yemi Okesanjo (2017). The policy gradient theorem provides a simplified form for the gradient, which can be used to derive on-policy and off-policy algorithms Ehsan Imani (2018). Natural policy gradients, which incorporate second-order information to improve convergence, form the foundation for state-of-the-art algorithms like Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) van Heeswijk (2022).

3.4 Methodology and Evaluation Metrics

In this paper, we will explore various RL algorithms, focusing on Q-learning and its variants, as well as policy gradient methods. We will delve into their theoretical foundations, convergence properties, and practical limitations. To assess the performance of these algorithms, we will use evaluation metrics such as cumulative reward, convergence speed, and sample efficiency. By comparing the performance of different algorithms, we aim to provide insights into their strengths and weaknesses, and identify potential areas for improvement and future research directions.

References

Adam Shwartz Arie Leizarowitz. Exact finite approximations of average-cost countable markov decision processes. *arXiv preprint arXiv:0711.2185*, 2007. URL http://arxiv.org/abs/0711.2185v1.

David Barber. Smoothed q-learning. *arXiv preprint arXiv:2303.08631*, 2023. URL http://arxiv.org/abs/2303.08631v1.

Martha White Ehsan Imani, Eric Graves. An off-policy policy gradient theorem using emphatic weightings. *arXiv preprint arXiv:1811.09013*, 2018. URL http://arxiv.org/abs/1811.09013v2.

Omri N. Solan Ehud Lehrer, Eilon Solan. The value functions of markov decision processes. *arXiv preprint arXiv:1511.02377*, 2015. URL http://arxiv.org/abs/1511.02377v1.

Miles Brundage Anil Anthony Bharath Kai Arulkumaran, Marc Peter Deisenroth. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*, 2017. URL http://arxiv.org/abs/1708.05866v2.

Vikram Krishnamurthy. Reinforcement learning: Stochastic approximation algorithms for markov decision processes. *arXiv preprint arXiv:1512.07669*, 2015. URL http://arxiv.org/abs/1512.07669v1.

Mattis Manfred Kmmerer. On policy gradients. *arXiv preprint arXiv:1911.04817*, 2019. URL http://arxiv.org/abs/1911.04817v1.

Morten Goodwin Paal Engelstad Li Meng, Anis Yazidi. Expert q-learning: Deep reinforcement learning with coarse state values from offline expert examples. *arXiv preprint arXiv:2106.14642*, 2021. URL http://arxiv.org/abs/2106.14642v3.

Agam Shah Mahipal Jadeja, Neelanshi Varia. Deep reinforcement learning for conversational ai. *arXiv preprint arXiv:1709.05067*, 2017. URL http://arxiv.org/abs/1709.05067v1.

Thomas Brihaye Paulin Fournier Nathalie Bertrand, Patricia Bouyer. Taming denumerable markov decision processes with decisiveness. *arXiv preprint arXiv:2008.10426*, 2020. URL http://arxiv.org/abs/2008.10426v1.

Kashu Yamazaki Khoa Luu Marios Savvides Ngan Le, Vidhiwar Singh Rathour. Deep reinforcement learning in computer vision: A comprehensive survey. *arXiv preprint arXiv:2108.11510*, 2021. URL http://arxiv.org/abs/2108.11510v1.

Billy Okal Philip S. Thomas. A notation for markov decision processes. *arXiv preprint arXiv:1512.09075*, 2015. URL http://arxiv.org/abs/1512.09075v2.

Shengqi Shen Jun Yang Meijing Zhao Kaiqi Huang Bin Liang Liang Wang Qiyue Yin, Tongtong Yu. Distributed deep reinforcement learning: A survey and a multi-player multi-agent learning toolbox. *arXiv preprint arXiv:2212.00253*, 2022. URL http://arxiv.org/abs/2212.00253v1.

Mattia Rigotti Rong Zhu. Self-correcting q-learning. *arXiv preprint arXiv:2012.01100*, 2020. URL http://arxiv.org/abs/2012.01100v2.

Alexander D'yakonov Sergey Ivanov. Modern deep reinforcement learning algorithms. *arXiv preprint arXiv:1906.10025*, 2019. URL http://arxiv.org/abs/1906.10025v2.

W. J. A. van Heeswijk. Natural policy gradients in reinforcement learning explained. *arXiv preprint arXiv:2209.01820*, 2022. URL http://arxiv.org/abs/2209.01820v1.

Zhipeng Wang Huijuang Wang Petter Holme Zi-Ke Zhang Xiu-Xiu Zhan, Chuang Liu. Measuring and utilizing temporal network dissimilarity. *arXiv preprint arXiv:2111.01334*, 2021. URL http://arxiv.org/abs/2111.01334v1.

Victor Kofia Yemi Okesanjo. Revisiting stochastic off-policy action-value gradients. *arXiv preprint arXiv:1703.02102*, 2017. URL http://arxiv.org/abs/1703.02102v2.